PLOS GENETICS

RESEARCH ARTICLE

A spectral theory for Wright’s inbreeding coefficients and related quantities

Olivier François*, Clément Gain

Université Grenoble-Alpes, Grenoble, France

* olivier.francois@univ-grenoble-alpes.fr

Abstract

Wright’s inbreeding coefficient, F_{ST}, is a fundamental measure in population genetics. Assuming a predefined population subdivision, this statistic is classically used to evaluate population structure at a given genomic locus. With large numbers of loci, unsupervised approaches such as principal component analysis (PCA) have, however, become prominent in recent analyses of population structure. In this study, we describe the relationships between Wright’s inbreeding coefficients and PCA for a model of K discrete populations. Our theory provides an equivalent definition of F_{ST} based on the decomposition of the genotype matrix into between and within-population matrices. The average value of Wright’s F_{ST} over all loci included in the genotype matrix can be obtained from the PCA of the between-population matrix. Assuming that a separation condition is fulfilled and for reasonably large data sets, this value of F_{ST} approximates the proportion of genetic variation explained by the first ($K - 1$) principal components accurately. The new definition of F_{ST} is useful for computing inbreeding coefficients from surrogate genotypes, for example, obtained after correction of experimental artifacts or after removing adaptive genetic variation associated with environmental variables. The relationships between inbreeding coefficients and the spectrum of the genotype matrix not only allow interpretations of PCA results in terms of population genetic concepts but extend those concepts to population genetic analyses accounting for temporal, geographical and environmental contexts.

Author summary

Principal component analysis (PCA) is the most-frequently used approach to describe population genetic structure from large population genomic data sets. In this study, we show that PCA not only estimates ancestries of sampled individuals, but also computes the average value of Wright’s inbreeding coefficient over the loci included in the genotype matrix. Our result shows that inbreeding coefficients and PCA eigenvalues provide equivalent descriptions of population structure. As a consequence, PCA extends the definition of those coefficients beyond the framework of allelic frequencies. We give examples on how F_{ST} can be computed from ancient DNA samples for which genotypes are corrected for coverage, and in an ecological genomic example where a proportion of genetic variation is explained by environmental variables.

Citation: François O, Gain C (2021) A spectral theory for Wright’s inbreeding coefficients and related quantities. PLoS Genet 17(7): e1009665. https://doi.org/10.1371/journal.pgen.1009665

Editor: Jonathan Marchini, Regeneron Genetics Center, UNITED STATES

Received: December 9, 2020

Accepted: June 13, 2021

Published: July 19, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting information files. All codes necessary to reproduce the simulations and data analyses presented in our study are in an R package available from the following link: https://github.com/bcm-uga/spectralfst.

Funding: The study was supported by the grant ANR-19-P3IA-0003 MIAI@Grenoble-Alpes – PEG: Predictive Ecological Genomics to OF from the Agence Nationale de la Recherche (https://anr.fr/) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Defined by Sewall Wright and Gustave Malécot, the fixation index or inbreeding coefficient, F_{ST}, measures the amount of genetic diversity found between populations relative to the amount within populations [1, 2]. Used as a measure of population differentiation, F_{ST} is among the most widely used descriptive statistics in population and evolutionary genetics [3–7]. Inbreeding coefficients were originally defined for the analysis of allelic frequencies at a single genetic locus. With the amount of data available to present-day or ancient population genomic studies, principal component analysis (PCA) and model-based estimation algorithms, such as STRUCTURE, have emerged as alternative ways to describe population structure from multilocus genotype matrices [8–12].

Assuming that the columns of the genotype matrix are either centered or scaled, PCA computes the eigenvalues and eigenvectors of the sample covariance matrix. The first eigenvectors—or axes—summarize the directions which account for most of the genetic variation, and the eigenvalues represent the variances of projected samples along the axes. Eigenvalues and eigenvectors can be computed efficiently by using the singular value decomposition (SVD) of the column-centered data matrix [13]. PCA has been considered very early in human biology, and has become a popular method to study the genetic structure of populations [14, 15]. Inference from PCA is justified from the fact that, similar to STRUCTURE, the projections of individuals on principal axes reveal their degree of admixture with source populations when these sources are represented in the sample [10, 16–18].

Although the relationships between PCA projections and admixture estimates are well-understood, difficulties of interpreting PCA eigenvalues still remain. The main contributions in that direction were restricted to models of divergence between two populations. The arguments were based on random matrix theory (RMT) [10, 19] and coalescent theory [16]. We note that connections between F_{ST} and PCA are not only important for description of population structure, but also in genome scans for selection where the distribution of PCA loadings can be used to detect regions with signature of divergent selection [20–23]. Based on RMT, Ref. [10] proposed a threshold value of F_{ST} for two populations with equal sample sizes. Below the threshold, there should be essentially no evidence of population structure. The coalescent approach relied on a relationship between F_{ST} and coalescent time for a pair of genes from a single subpopulation and that of a pair of genes from the collection of subpopulations [6]. For a model of divergence between two populations, theoretical results for coalescent times were used to demonstrate a link between the leading eigenvalue of PCA and F_{ST} [16]. Results in Ref. [16] might be extended to simple models of population structure with explicit formulas for coalescent times [24], but the results are not straightforward. While coalescent theory and RMT have provided relationships between F_{ST} and PCA in simple cases, the general conditions under which they are valid and their extensions to more than two populations are unknown.

In this study, we develop a spectral theory of genotype matrices to investigate the relationships between PCA and Wright's coefficients in discrete population models. Our theoretical framework assumes that the observed genotypes correspond to the sampling of K discrete populations. Decomposing the genotype matrix as a sum of between and within-population matrices, we extend the results obtained in [10, 16, 19, 25]. Our main result states that the mean value of F_{ST} over loci is equal to the squared Hilbert-Schmidt norm of the between-population matrix, which can be computed by a spectral analysis. Under a separation condition bearing on the between and within-population matrices, the sum of the first $(K - 1)$ eigenvalues of scaled PCA approximates the mean value of F_{ST} over loci. To describe residual variation not explained by the discrete population model, we rely on approximations of the eigenvalues of the within-population matrix from RMT [10, 26].

A corollary of the theory is an alternative definition of inbreeding coefficients that allows us to extend F_{ST} to adjusted or *surrogate* genotypes, such as genotype likelihoods and other modifications of allele counts [27]. To illustrate the new definition, we compute F_{ST} for ancient human DNA samples after performing correction for genomic coverage and for distortions due to difference in sample ages [28]. In a second application, we compute F_{ST} for Scandinavian samples of *Arabidopsis thaliana* after removing genetic variation associated with environmental variables taken from a climate database [29, 30].

Results and discussion

Partitioning of genetic variation

Consider a sample of n unrelated individuals for which a large number of loci are genotyped, resulting in a matrix, $\mathbf{X} = (x_{i\ell})$, with n rows and L columns. For haploids, we set $x_{i\ell} = 0, 1$, and for diploids $x_{i\ell} = 0, 1, 2$ to count the number of derived alleles at locus ℓ for individual i. Dealing with autosomes, we simplify our presentation by considering a sample of diploids as being represented by a sample of haploids having twice the original sample size. For unphased data, we take a random phase. Although not a necessary condition, the loci are assumed to be unlinked, or obtained after a linkage disequilibrium (LD)-pruning algorithm applied to the genotype matrix [20, 31]. We use the term locus as a shorthand for single-nucleotide polymorphism (SNP), although most of our analyses could include non-polymorphic sites. Following Wright's approach to the description of population structure, our main assumption is that individuals are sampled from K predefined discrete populations. Examples of discrete population models underlying our assumptions include Wright's island models, coalescent models of divergence and F-models [6, 32, 33]. Application to F-models will be described afterwards.

To analyze population structure, PCA can be performed after centering and sometimes after scaling the genotype matrix. The scaled matrix is denoted by $\mathbf{Z}^{\mathrm{sc}}$ and the unscaled centered matrix is denoted by $\mathbf{Z}$. Scaled PCA computes the eigenvalues, $\rho_k^2(\mathbf{Z}^{\mathrm{sc}})$, of the empirical correlation matrix. Unscaled PCA computes the eigenvalues, $\sigma_k^2(\mathbf{Z})$, of the empirical covariance matrix [9, 26]. The eigenvalues are ranked in decreasing order, and $\rho_k^2(\mathbf{Z}^{\mathrm{sc}})/L$ is usually interpreted as the proportion of variance explained by the kth axis of the PCA. PCA can be performed via the SVD algorithm. In this case, the eigenvalues of scaled (or unscaled) PCA correspond to the squared singular values of the scaled (or centered) matrix divided by $\sqrt{n}$ [9, 26].

To establish relationships between PCA and inbreeding coefficients, we decompose the centered matrix into a sum of two matrices, $\mathbf{Z} = \mathbf{Z}_{\mathrm{ST}} + \mathbf{Z}_{\mathrm{S}}$, corresponding to between and within-population components. The decomposition is performed as follows. Let i be an individual sampled from population k. At a particular locus, ℓ, the genotype, $x_{i\ell}$, is equal 0 or 1 (derived allele), and $p_{k\ell}$ denotes the derived allele frequency in population k at this locus. The coefficient of the centered matrix, $z_{i\ell}$, is equal to $z_{i\ell} = \sum_{j\neq k} c_j(p_{k\ell} - p_{j\ell}) + (x_{i\ell} - p_{k\ell})$, where $c_k = n_k/n$ represents the proportion of individuals sampled from population k. In this formulation, the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}$, has general term $z^{st}_{i\ell} = \sum_{j\neq k} c_j(p_{k\ell} - p_{j\ell})$, repeated for all individuals in population k. By construction, the rank of $\mathbf{Z}_{\mathrm{ST}}$ is equal to $(K - 1)$. The within-population matrix, $\mathbf{Z}_{\mathrm{S}}$, has general term $z^{s}_{i\ell} = x_{i\ell} - p_{k\ell}$. A very similar decomposition holds for the scaled matrix, defined as $z^{sc}_{i\ell} = z_{i\ell}/\sqrt{P_\ell(1 - P_\ell)}$, where $P_\ell = \sum_{k=1}^{K} c_k p_{k\ell}$ is the derived allele frequency in the total sample at locus ℓ (see Box 1 for the notations).

Box 1. Notations

n Haploid sample size

L Number of genomic loci

F_{ST} Wright's fixation index, computed from Nei's formula with correction for unequal sample sizes

H_{S} Within-population genetic diversity

H_{T} Genetic diversity in the total population

D_{ST} Among (or between) population genetic diversity

$\mathbf{X}$ Matrix of SNP genotypes for n individuals at L loci

$\mathbf{P}$ Vector of SNP frequency for the L loci

$\mathbf{Z}$ Matrix of centered genotypes, $\mathbf{X} - \mathbf{P}$

$\mathbf{Z}^{\mathrm{sc}}$ Matrix of scaled genotypes, $\mathbf{Z}/\sqrt{\mathbf{P}(1-\mathbf{P})}$

$\mathbf{Z}_{\mathrm{ST}}$ An $n \times L$ matrix describing between-population data repeated for individuals from a same population

$\mathbf{Z}_{\mathrm{S}}$ An $n \times L$ matrix describing within-population data

$\sigma_k^2(\mathbf{Z})$ Eigenvalues of the empirical covariance matrix (unscaled PCA)

$\rho_k^2(\mathbf{Z}^{\mathrm{sc}})$ Eigenvalues of the empirical correlation matrix (scaled PCA), also equal to L times the proportions of variance explained by the principal axes

Spectral analysis of inbreeding coefficients

Consider n samples from K discrete populations and define D_{ST} and F_{ST} according to Wright's [1] and Nei's [4, 34] equations, allowing for unequal population sample sizes. At a particular locus, set $H_{\mathrm{S}} = 2\sum_{k=1}^{K} c_k p_k(1-p_k)$ and $H_{\mathrm{T}} = 2P(1 - P)$, then we have $D_{\mathrm{ST}} = H_{\mathrm{T}} - H_{\mathrm{S}}$. Wright's inbreeding coefficient is defined as

$$F_{\mathrm{ST}} = D_{\mathrm{ST}}/H_{\mathrm{T}}\,.$$

Our main result states that the mean value of F_{ST} across loci can be computed from the singular values of the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}$. A similar relationship is also established for D_{ST} and for the unscaled matrix, $\mathbf{Z}_{\mathrm{ST}}$. The singular values of the between and within-population matrices can be evaluated from the SVD algorithm. The computational cost of those operations is equal to the computational cost of the PCA of the genotype matrix, of order $O(n^2 L)$. This cost could be reduced by using various methods, for example by computing the first $K - 1$ singular values only. All conclusions below are valid regardless of any population genetic model. The results also remain valid when genotypes are conditioned on having minor allele frequency greater than a given threshold, and when the loci are physically linked or when they exhibit LD. We use the notation $\mathbb{E}[Q] = \sum_{\ell=1}^{L} Q_\ell/L$ to denote the average value of the quantity Q_ℓ across loci.

Theorem 1. *Let $K \geq 2$. Let $\mathbf{Z}$ and $\mathbf{Z}^{\mathrm{sc}}$ be the unscaled and scaled genotype matrix respectively. Define $\mathbf{Z}_{\mathrm{ST}}$ and $\mathbf{Z}^{\mathrm{sc}}_{\mathrm{ST}}$ as in the previous section. We have*

$$\mathbb{E}[F_{\mathrm{ST}}] = \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}^{\mathrm{sc}}_{\mathrm{ST}})/L\,, \tag{1}$$

and

$$\mathbb{E}[D_{\mathrm{ST}}]/2 = \sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})/L\,. \tag{2}$$

The key arguments for those results involve matrix norms, and they can be found in S1 Text. By the Pythagorean theorem, we have $\|\mathbf{Z}\|^2 = \|\mathbf{Z}_{\mathrm{ST}}\|^2 + \|\mathbf{Z}_{\mathrm{S}}\|^2$ (S1 Text). According to this result, Theorem 1 can be reformulated using within-population matrices as follows.

Corollary. *Let $K \geq 2$. Let $\mathbf{Z}$ and $\mathbf{Z}^{\mathrm{sc}}$ be the unscaled and scaled genotype matrix respectively. Define $\mathbf{Z}_{\mathrm{S}}$ and $\mathbf{Z}^{\mathrm{sc}}_{\mathrm{S}}$ as in the previous section. We have*

$$\mathbb{E}[F_{\mathrm{ST}}] = 1 - \sum_{j=1}^{n-K} \rho_j^2(\mathbf{Z}^{\mathrm{sc}}_{\mathrm{S}})/L\,. \tag{3}$$

and

$$\mathbb{E}[D_{\mathrm{ST}}] = \mathbb{E}[H_{\mathrm{T}}] - 2\sum_{j=1}^{n-K} \sigma_j^2(\mathbf{Z}_{\mathrm{S}})/L\,. \tag{4}$$

Inbreeding coefficients for surrogate genotypes

Besides an interest in connecting population genetic theory to the spectrum of the genotype matrix, the results in Eqs (1) and (2) have important consequences for data analysis. First, the results support the definition of F_{ST} for multilocus genotypes as an average of ratios rather than a ratio of averages [35, 36]. More importantly, Theorem 1 leads to alternative definitions of Nei's D_{ST} and Wright's F_{ST} from population genetic data. As will be demonstrated by applications to ancient DNA and to ecological genomics, the main interest in those new definitions is their straightforward extension to adjusted genotypic matrices, providing statistics analogous to F_{ST} for modified genotypic values. For example, adjusted genotypic matrices arise when correcting for biases due to technical artifacts, including batch effects and genomic coverage in population genomic data [37, 38]. In general, F_{ST} could be adjusted for any specific effect by considering the residuals of latent factor regression models [39–41]. More specifically, for $\mathbf{Z}$ (or $\mathbf{Z}^{\mathrm{sc}}$) and for a set of measured covariates, $\mathbf{Y}$, latent factor regression models estimate a matrix of surrogate genotypes, $\mathbf{W}$, by adjusting a regression model of the form

$$\mathbf{Z} = \mathbf{Y}\mathbf{B}^T + \mathbf{W} + \epsilon\,.$$

In this model, the $\mathbf{B}$ matrix contains effect sizes for each variable in $\mathbf{Y}$, and ϵ is a matrix that represents centered errors. The latent matrix $\mathbf{W}$ has a specified rank, k, lower than n minus the number of covariates. The rank k corresponds to the number of latent factors incorporated in the model. The matrix $\mathbf{Z}^{\mathrm{adj}} = \mathbf{W} + \epsilon$ leads to a definition of an adjusted inbreeding coefficient, $F^{\mathrm{adj}}_{\mathrm{ST}}$. The adjusted inbreeding coefficient can be computed as the squared norm of the between-population matrix, $\mathbf{Z}^{\mathrm{adj+sc}}_{\mathrm{ST}}$, after scaling. Note that this definition considers quantitative values observed at each locus, and is similar to a population genetic quantity called Q_{ST} [42, 43].

Alternatively, $F_{\mathrm{ST}}^{\mathrm{adj}}$ can be computed from the average coefficient of determination, R^2, obtained from the regression of each scaled surrogate genotype on the population labels. The definitions are equivalent, and we have

$$\mathbb{E}[F_{\mathrm{ST}}^{\mathrm{adj}}] = \sum_{k=1}^{K-1} \rho^2(\mathbf{Z}_{\mathrm{ST}}^{\mathrm{adj+sc}})/L = \mathbb{E}[R^2]\,.$$

Inbreeding coefficients and PCA eigenvalues

Having established that the mean value of F_{ST} across loci can be computed from the leading eigenvalues of the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}$, the next question is to ask whether similar results hold for the leading eigenvalues of the PCA of the scaled matrix,

$$\mathbb{E}[F_{\mathrm{ST}}] \approx \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}^{\mathrm{sc}})/L\,, \tag{5}$$

and for the PCA of the centered matrix,

$$\mathbb{E}[D_{\mathrm{ST}}] \approx 2\sum_{k=1}^{K-1} \sigma_k^2(\mathbf{Z})/L\,. \tag{6}$$

Those results require that the ranked eigenvalues of $\mathbf{Z}/\sqrt{n}$ sort into approximate eigenvalues of $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$ followed by approximate eigenvalues of $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$. Said differently, the $(K - 1)$th singular value of $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$ must separate from the leading singular value of $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$

$$\sigma_{K-1}^2(\mathbf{Z}_{\mathrm{ST}}) > \sigma_1^2(\mathbf{Z}_{\mathrm{S}})\,. \tag{7}$$

We suppose that the ratio L/n is constant for large L and n, and make the following assumptions: 1) The separation condition Eq (7) is verified, 2) The leading eigenvalue of $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$ is of order $(1/\sqrt{n} + 1/\sqrt{L})^2$ (RMT hypothesis). Then, under those conditions, the accuracy of the approximation in Eqs (5) and (6) is of order $O(K/L)$. More precisely, for any singular value, $\sigma_{\mathrm{k}}(\mathbf{Z}_{\mathrm{ST}})$, of $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$, there exists a singular value, $\sigma_{\mathrm{k}}(\mathbf{Z})$, of $\mathbf{Z}/\sqrt{n}$ such that we have

$$|\sigma_k^2(\mathbf{Z})/L - \sigma_k^2(\mathbf{Z}_{\mathrm{ST}})/L| = O(1/L)\,, \quad k = 1, \ldots, K-1\,.$$

A similar result holds for the first $K - 1$ eigenvalues of scaled matrices, $\rho_k^2(\mathbf{Z}^{\mathrm{sc}})$ and $\rho_k^2(\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}})$. In other words, the mean value of F_{ST} across loci can be approximated from the sum of the $(K - 1)$ leading eigenvalues of the PCA with an accuracy proportional to the number of populations and to the inverse of the number of unlinked loci in the genotype matrix. Mathematical arguments for those results are detailed in S1 Text.

Poor approximations may be caused by insufficient sample size, incorrect definition of populations, inclusion of individuals with mixed ancestry, spatial structure, etc. Poor approximations may also be accompanied by failure to verify the separation condition Eq (7), as our simulations will illustrate afterwards. In addition, the results show that F_{ST} and PCA exhibit similar biases, for example, when the sampling design is uneven or when loci are filtered out of the genotype matrix [16, 31, 44]. We note that the RMT hypothesis for the residual matrix, $\mathbf{Z}_{\mathrm{S}}$, is difficult to prove for population genetic models. Like [10], we will rely on simulations to show that RMT describes residual variation in population genetic models accurately. In empirical data analyses, checking the residual matrix for agreement with RMT will also provide an informal test for the number of components in PCA similar to Cattel's elbow rule [45].

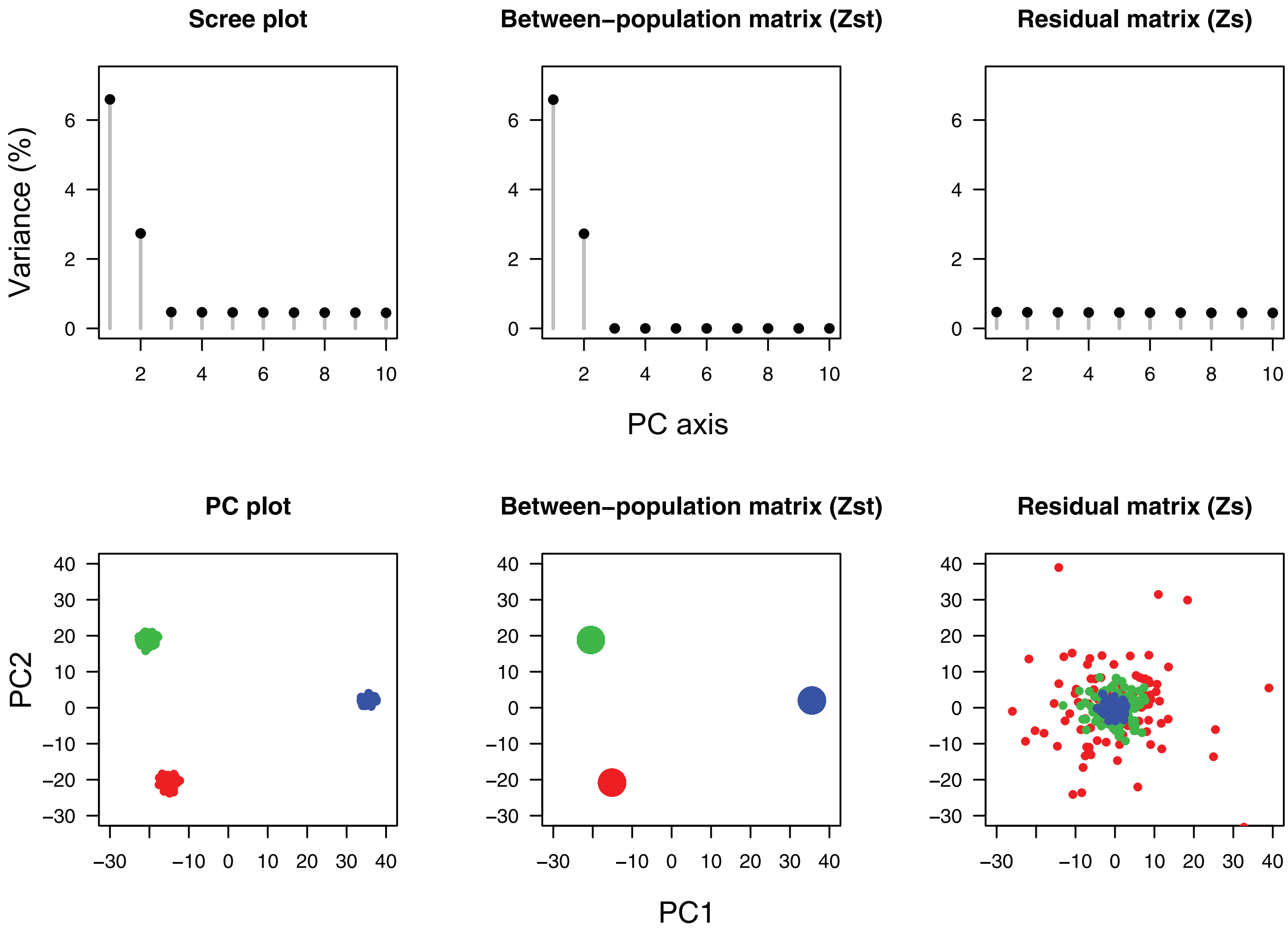

Fig 1. Spectral analysis of a three-population model. PCA scree plots and PC plots for the scaled matrix, $\mathbf{Z}^{\mathrm{sc}}$, for the between-population matrix, $\mathbf{Z}^{\mathrm{sc}}_{\mathrm{ST}}$, and for the residual matrix, $\mathbf{Z}^{\mathrm{sc}}_{\mathrm{S}}$ of simulated data. The simulation was performed for $n = 300$ individuals and an F-model ($F_1 = 5\%$, $F_2 = 10\%$, $F_3 = 30\%$) with ancestral frequency drawn from a beta(1,4) distribution.

https://doi.org/10.1371/journal.pgen.1009665.g001

Brief example

To illustrate the approximation of $\mathbb{E}[F_{\mathrm{ST}}]$ by the leading eigenvalues of the PCA, we present a short simulation example, in which a genotype matrix was generated according to a three-population F-model. Simulation studies of F-models and additional examples based on real data will be developed more extensively later on. In this first simulation example, the average ancestral frequency was equal to 20%, and the drift parameters for the three populations were equal to $F_1 = 5\%$, $F_2 = 10\%$ and $F_3 = 30\%$. Populations 1 and 2 were genetically closer to each other than to population 3, which was the most diverged population. Three hundred samples ($n_k = 100$, for $k = 1, 2, 3$) were genotyped at 10,000 loci. PCA was conducted on $L = 9740$ SNP loci after monomorphic loci were removed. The average value of F_{ST} across loci was equal to 9.52%, and approximated the sum of the leading eigenvalues of the PCA (9.54%) accurately. The first axes of the PCA explained 6.78%, 2.76%, and 0.47% of the total variation respectively (Fig 1). The non-null eigenvalues of $\mathbf{Z}^{\mathrm{sc}}_{\mathrm{ST}}/\sqrt{n}$, 6.77% and 2.75%, were close to the values obtained for the first two PCs. As stated in Theorem 1, their sum was

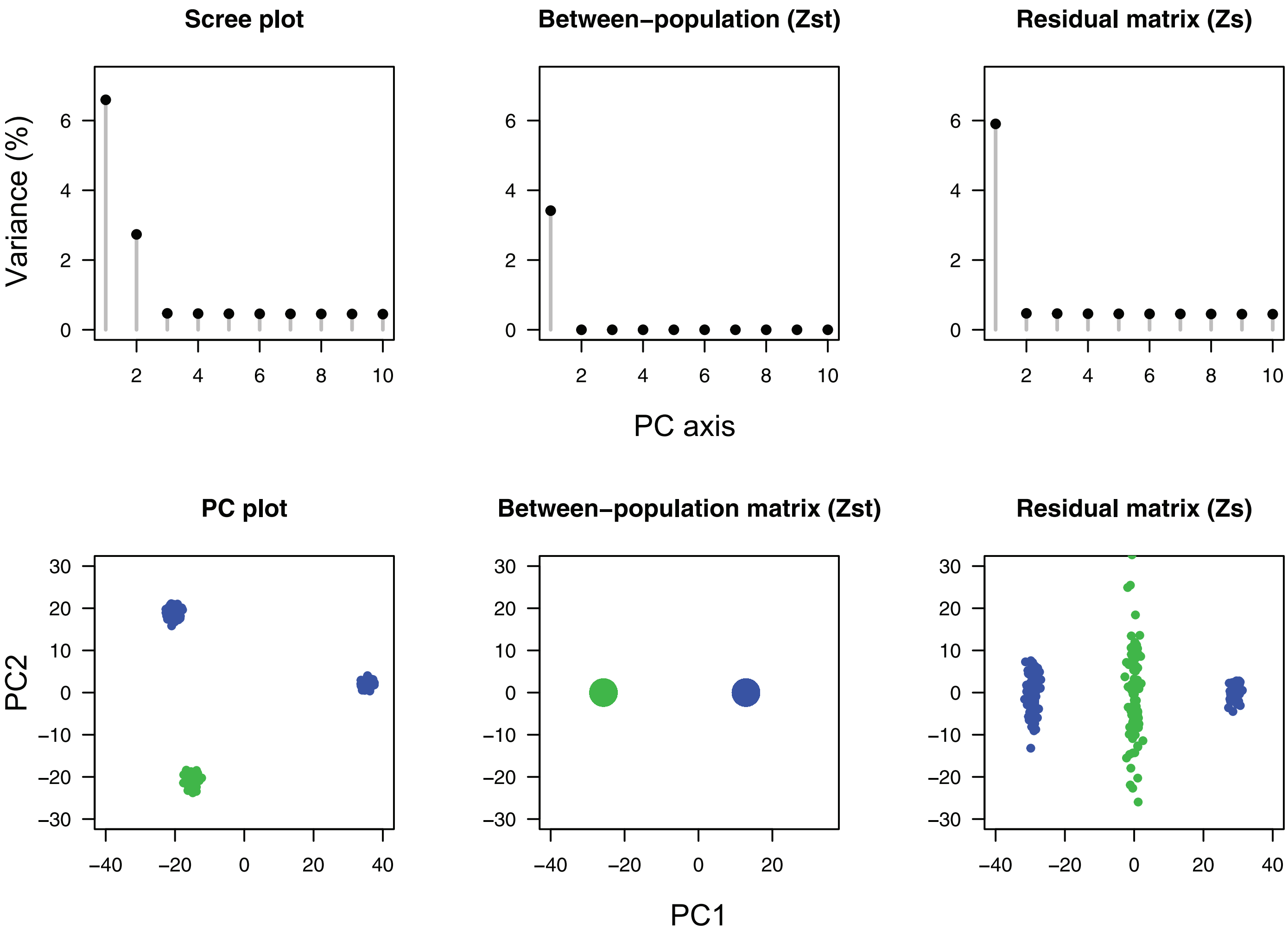

Fig 2. Spectral analysis with incorrectly labelled population samples ($K = 2$). For the same genotype matrix as in Fig 1, samples from populations 2 and 3 (blue) were grouped against population 1 (green). PCA scree plots and PC plots for the scaled matrix, $\mathbf{Z}^{\text{sc}}$, for the between-population matrix, $\mathbf{Z}^{\text{sc}}_{\text{ST}}$, and for the residual matrix, $\mathbf{Z}^{\text{sc}}_{\text{S}}$ of simulated data. F_{ST} was lower than the leading eigenvalue of the residual matrix, and it differed from the first PC eigenvalue.

https://doi.org/10.1371/journal.pgen.1009665.g002

equal to $\mathbb{E}[F_{\text{ST}}]$. Clearly, the smallest eigenvalue of $\mathbf{Z}^{\text{sc}}_{\text{ST}}/\sqrt{n}$ separated from the leading value of $\mathbf{Z}^{\text{sc}}_{\text{S}}/\sqrt{n}$ (0.47%), which was close to the eigenvalue for the third PC and to its prediction from RMT (0.31%, Fig 1).

To show the effect of having incorrectly labelled population samples, we considered the same genotype matrix, and replicated the analyses after grouping the "paraphyletic" samples from populations 2 and 3 against the least diverged sample from population 1 (Fig 2, $n_1 = 200$ and $n_2 = 100$, $K = 2$). The average F_{ST} value was equal to 3.46%, and failed to approximate the first eigenvalue of the PCA (6.78%). The leading value of $\mathbf{Z}^{\text{sc}}_{\text{S}}/\sqrt{n}$ (6.06%) did not verify the separation condition, and it differed from its RMT prediction (0.32%). The PC plot for $\mathbf{Z}^{\text{sc}}_{\text{S}}$ provided evidence of residual population structure within the paraphyletic population sample. Like for a regression analysis, those results outlined the usefulness of visualizing the residual matrix in order to evaluate the number of populations from the genotype matrix (Figs 1 and 2).

Single population models

In a first series of simulations without population structure, we investigated whether RMT predictions were valid for F-models. For single population F-models, the results supported that the leading eigenvalues of PCA were accurately predicted by the Marchenko-Pastur distribution (S1 Fig). Then we investigated whether condition Eq (7) could be verified when there was no structure in the data, and two population samples were wrongly defined from a preliminary structure analysis. We ran two-hundred simulations of single population models ($n = 100$ and $L \approx 10,000$), and, for each data set, we partitioned the samples in two groups according to the sign of their first principal component. This procedure maximized the likelihood of detecting artificial groups, leading to an average $F_{\mathrm{ST}} \approx 1.1\%$. For those artificial groups, we computed the non-null singular value of the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}^{\mathrm{sc}}/\sqrt{n-1}$, and the leading singular value of the within-population matrix, $\mathbf{Z}_{\mathrm{S}}^{\mathrm{sc}}/\sqrt{n-1}$. For the simulations, the separation condition was never verified, rejecting population structure in all cases (S2(A) Fig). For smaller sample sizes ($n = 10$ and $L \approx 1,000$), the separation condition was erroneously checked in 21% simulations, indicating that we had less power to discriminate among artificial groups with small sample sizes (S2(B) Fig). Those results were also consistent with difficulties reported for between-group PCA [46].

Two-population models

To check whether the expected values of F_{ST} and D_{ST} were approximated from the first eigenvalues of PCA, we performed simulations of F-models with two populations. For these simulations, the separation condition was verified in all data sets. There was an almost perfect fit of the leading eigenvalue for centered PCA, $\sigma_1^2(\mathbf{Z})$, with the average value of $D_{\mathrm{ST}}/2$ across loci and with its theoretical value in F-models (Fig 3A and S3 Fig, and S1 Text). There was also an almost perfect fit of the leading eigenvalue of scaled PCA, $\rho_1^2(\mathbf{Z}^{\mathrm{sc}})$, with the average value of F_{ST} across loci (Fig 3C). The second largest eigenvalues were accurately predicted by RMT both for unscaled and for scaled PCA (Fig 3B–3D). To detail those results for particular values of drift coefficients, we performed additional simulations for $F_1 = F_2 = 7\%$, also investigating the distribution of eigenvalues of the residual matrix (Fig 4). In a sample of $n = 200$ individuals and $L \approx 85,500$ SNPs, the first PC axis explained 3.11% of total genetic variation, corresponding to the average of F_{ST} across loci (3.11%, Fig 4A). The separation of between and within-population components was verified, and the second eigenvalue (0.536%) was very close to its prediction from RMT, given by $(1-\rho_1^2(\mathbf{Z}^{\mathrm{sc}})) \times (1/\sqrt{L} + 1/\sqrt{n-2})^2 = 0.537\%$ (Fig 4A). The distribution of residual eigenvalues, corresponding to within-population variation, was accurately modelled by the Marchenko-Pastur probability density function (Fig 4B). With a smaller sample of $n = 20$ individuals and $L \approx 12,500$ SNPs, the leading axis explained 5.24% of the total genetic variation, still matching the value of F_{ST} across loci (5.23%, Fig 4C). The Marchenko-Pastur density remained an accurate approximation to the bulk spectrum of residual eigenvalues (Fig 4D).

In another series of simulations with $F_1 = F_2 = 2\%$, we evaluated whether the sum of the leading eigenvalues for scaled PCA was close to the average of F_{ST} across loci when the separation condition was verified (S4 Fig). For small sample sizes ($n = 20$ and $L = 100$), the separation condition was not verified and the approximation was poor. For intermediate sample sizes ($n = 60$ and $L = 1000$), the approximation was more accurate when the separation condition was verified than when it was not. For large sample sizes ($n = 100$ and $L = 10000$), the separation condition was always verified and the approximation was accurate. Although the L/n ratio

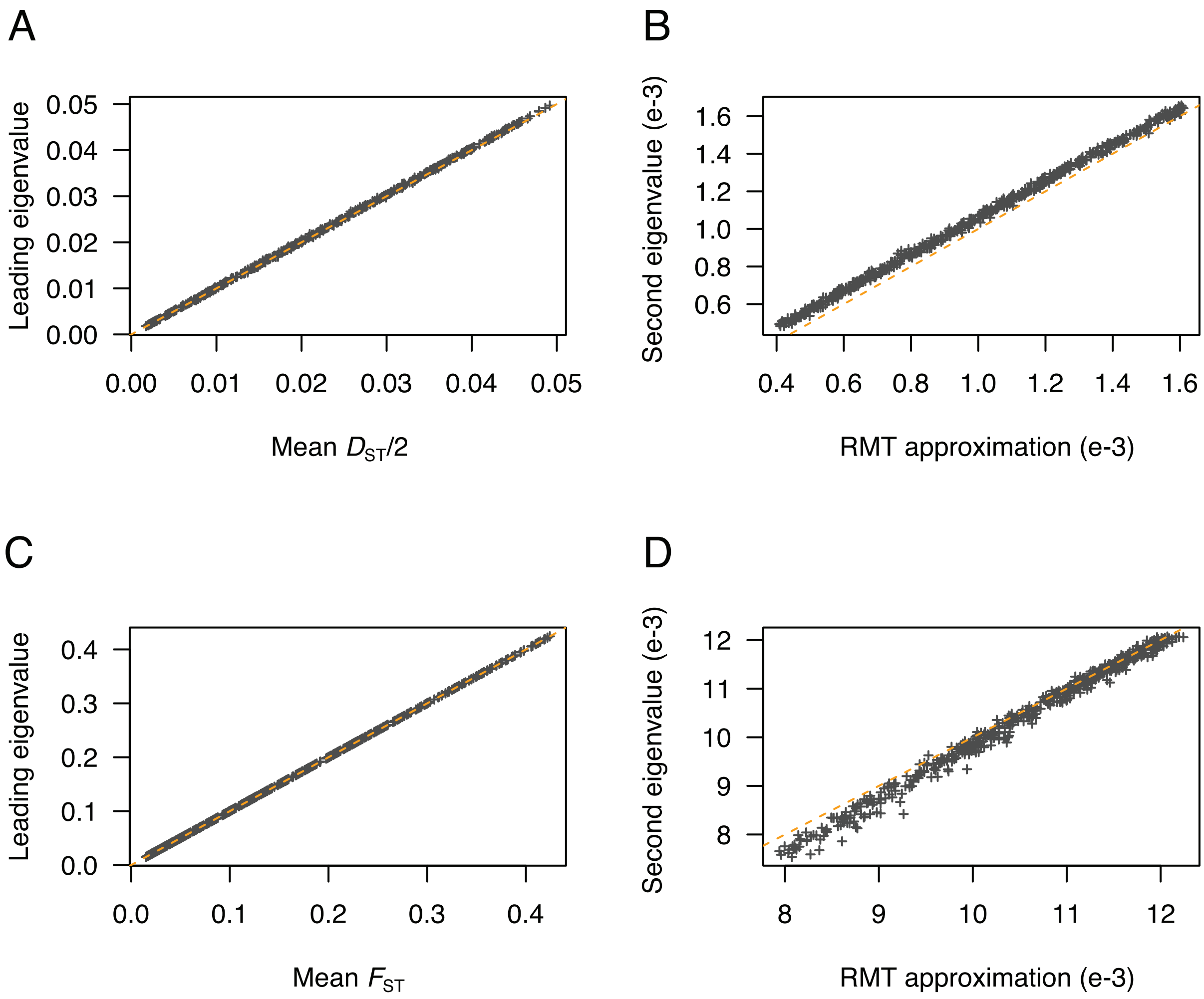

Fig 3. Comparison of D_{ST} and F_{ST} estimates with the leading PCA eigenvalues in two-population models. (**A**) Leading eigenvalues of centered PCA as a function of the mean of $D_{ST}/2$ across loci. (**B**) Second eigenvalue of centered PCA as a function of its approximation from RMT. (**C**) Leading eigenvalues of scaled PCA as a function of the mean of F_{ST} across loci. (**D**) Second eigenvalue of scaled PCA as a function of its approximation from RMT, which is given by $(1-\rho_1^2)\times(1/\sqrt{L}+1/\sqrt{n-2})^2$ (approximation of the largest eigenvalue of the residual matrix). The dashed lines correspond to the diagonal $y = x$. Simulations of F-models were performed for $n = 100$ individuals (inbreeding coefficients between 1% and 75%, first population sample proportion between 10% and 50%, ancestral frequency was drawn from a beta(1,4) distribution).

https://doi.org/10.1371/journal.pgen.1009665.g003

was not kept to a constant value, the accuracy of the approximation agreed with a predicted order of $O(1/L)$ (S5 Fig).

Next, we studied the relationship between leading eigenvalues and sample size, for $L = 100$ loci and $L = 100,000$ loci (S6 Fig). For smaller number of loci ($L = 100$) and smaller samples ($n \leq 80$), the data failed to verify the separation condition in some simulations. In those cases, population structure was not correctly evaluated by $\mathbb{E}[F_{ST}]$. The separation condition was verified in about 35% cases for $n = 10$ and in about 95% cases for $n \approx 80$. For the larger number of loci ($L \geq 100000$) or for larger sample sizes ($n \geq 100$), the separation condition was verified in all cases, and the leading eigenvalue converged to the theoretical value of $\mathbb{E}[F_{ST}]$ for infinitely large sample sizes. As for between-group PCA, the results suggest exaggerated differences among groups when sample sizes are very small relative to the number of loci [47].

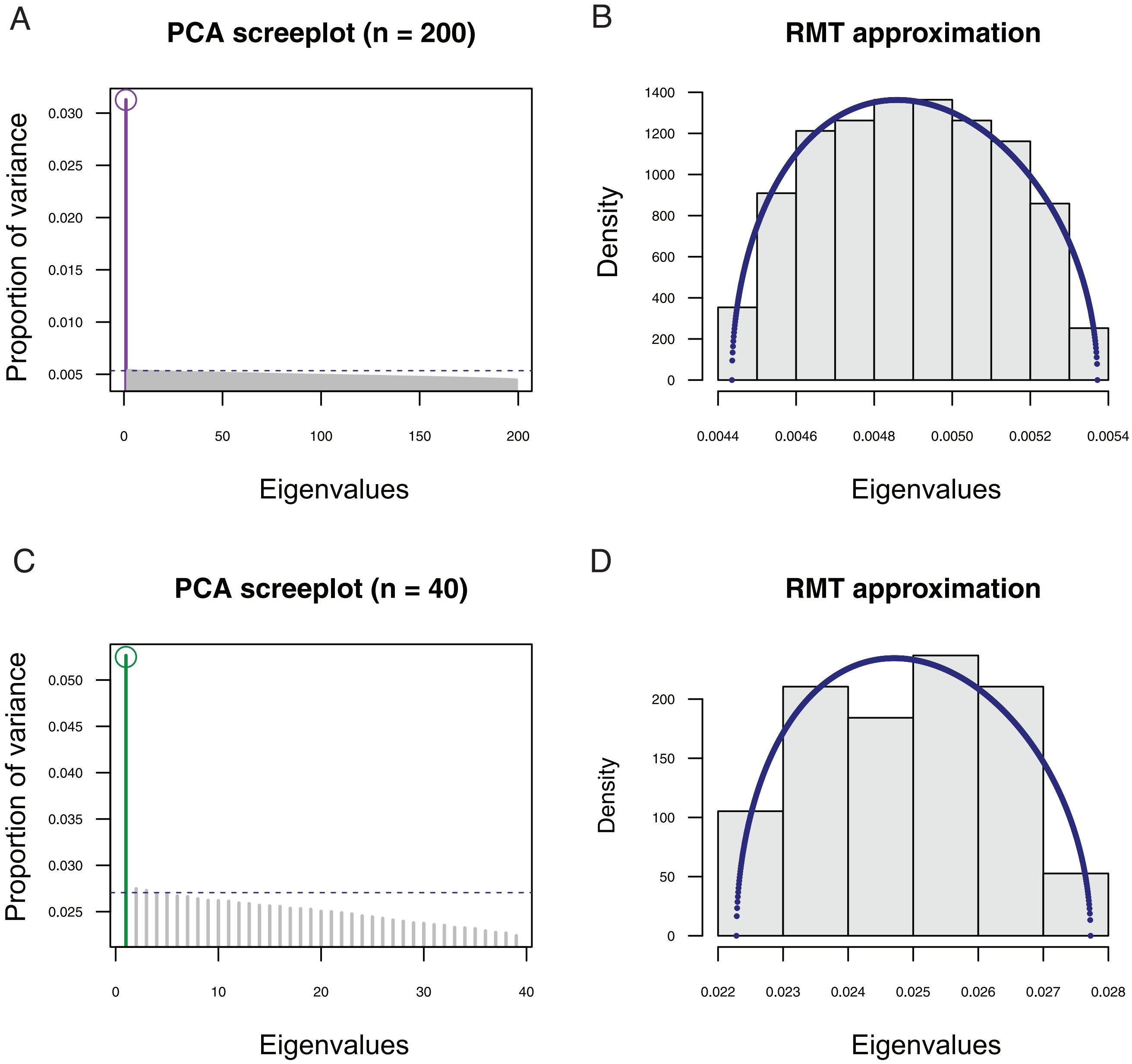

Fig 4. Scree plots and RMT approximations in two-population models. (**A**) Proportion of variance (eigenvalues) explained by PC axes, with a circle symbol representing the mean of F_{ST} across loci for $n = 200$ individuals and $L = 85,540$ SNPs. (**B**) Histogram of eigenvalues of the residual matrix, $\mathbf{Z}_{\mathrm{S}}/\sqrt{n-2}$, for the data simulated in A. (**C**) Proportion of variance for $n = 40$ individuals and $L = 12,650$ SNPs. (**D**) Histogram of eigenvalues of the residual matrix for the data simulated in C. The dashed lines in PCA scree-plots represent the RMT approximation of the leading eigenvalue of the residual matrix. The blue curve represents the Marchenko-Pastur probability density. Simulations of F-models were performed with p_{anc} drawn from a beta(1,9) distribution and $F_1 = F_2 = 7\%$.

https://doi.org/10.1371/journal.pgen.1009665.g004

More than three-population models

For F-models, the eigenvalues of the theoretical covariance matrix were analysed formally for small numbers of populations (S1 Text). To achieve this goal, we considered the covariance matrix of the random vector $\mathbf{z}$ defined by $\mathbf{z}_k = \sqrt{c_k}(p_k - P)$, for all $k = 1$ to K. The $K \times K$ covariance matrix of the random vector $\mathbf{z}$ could be obtained from the drift statistics $\mathcal{F}_2$ and $\mathcal{F}_3$ defined in [48, 49]. The coefficients of this matrix are

$\mathbf{\Lambda}_{j,k} = \sqrt{c_j c_k}\,\mathbb{E}[(p_j - P)(p_k - P)] = \sqrt{c_j c_k}\,\mathcal{F}_3(P; p_j, p_k)$, for $j \neq k$, and $\mathbf{\Lambda}_{k,k} = c_k \mathbb{E}[(p_k - P)^2] = c_k \mathcal{F}_2(P; p_k)$ otherwise. For $K = 3$, the eigenvalues of the $\mathbf{\Lambda}$ matrix were computed exactly (S1 Text).

We performed simulations of three-population F-models to check whether the data agreed with theoretical predictions for the leading eigenvalues, λ_1, and for D_{ST} and F_{ST}. With random drift coefficients ($n = 100$, $L = 20000$), the separation condition was verified in all simulated data sets. An almost perfect agreement between $\lambda_1 + \lambda_2$ and the mean value of $D_{\mathrm{ST}}/2$ (unscaled PCA) or F_{ST} (scaled PCA) was observed (S7(A)–S7(C) Fig). The leading eigenvalue of unscaled PCA exhibited a small but visible bias with respect to the value predicted for λ_1 (S7(B) Fig). The third eigenvalue of scaled PCA was close to the approximation provided by RMT (S7(D) Fig). To study cases in which the separation condition was not verified, we considered smaller number of genotypes ($L \leq 1000$) and lower values of drift coefficients ($F_k \leq 10\%$). For small values of n and L, a significant proportion of simulated data sets did not verify the separation condition (S8 Fig), although models were correctly specified. Those results provided additional evidence of biases in analyses of population structure with small data sets. Extending our simulation study to larger number of populations, we checked that the accuracy of the approximation of $\mathbb{E}[F_{\mathrm{ST}}]$ by the sum of the first $(K - 1)$ PCA eigenvalues was proportional to K/L when the separation condition was verified (S9 Fig). Poorer accuracy was observed when the data failed to verify the separation condition.

Human data

To provide evidence that the relationships between PCA eigenvalues and F_{ST} are verified for real genotypes, we computed F_{ST}, their approximation by PCA, and the leading eigenvalues of the residual matrix for pairs and triplets of human population samples from The 1000 Genomes Project [50] (Table 1 and S1 Table). In pairwise analyses excluding admixed samples, the separation condition was verified in all analyses at the exception of the pair CEU-IBS, formed of two closely related European samples. The leading eigenvalue of scaled PCA was accurately approximated by $\mathbb{E}[F_{\mathrm{ST}}]$, and the leading eigenvalue of the residual matrix was accurately predicted by RMT. For triplet analyses without admixed samples, the separation condition was also verified and the sum of the first two eigenvalues of scaled PCA was accurately approximated by $\mathbb{E}[F_{\mathrm{ST}}]$. RMT still predicted the leading eigenvalue of the residual matrix accurately. In pair and triplet analyses including admixed samples, the separation condition

Table 1. F_{ST} estimates for populations from The 1,000 Genomes Project.

	Lead. eigen. of PCA*	**F_{ST} across loci**	**Lead. eigen. res. matrix****	**RMT approximation*****
CHB-CEU	5.65%	5.65%	0.42%	0.48%
CHB-YRI	8.35%	8.35%	0.36%	0.37%
CEU-YRI	7.21%	7.21%	0.35%	0.37%
CEU-IBS	0.41%	0.38%	0.37%	0.41%
CEU-YRI-CHB	9.99%	9.98%	0.24%	0.26%
CEU-ASW	4.87%	4.55%	0.75%	0.52%
CEU-YRI-ASW	6.12%	5.82%	0.44%	0.29%

* Sum of the leading eigenvalues of the PCA

** Sum of the leading eigenvalues of the within-population (residual) matrix

*** RMT approximation for the leading eigenvalue of the within-population matrix

IBS: Iberian ($n = 147$), **CHB**: Han Chinese in Beijing ($n = 100$), **YRI**: Yoruba ($n = 158$), **CEU**: Utah residents with European ancestry ($n = 104$). **ASW**: Americans of African Ancestry in SW USA ($n = 97$).

https://doi.org/10.1371/journal.pgen.1009665.t001

was fulfilled in all analyses at the exception of the pair ACB-ASW (S1 Table). The approximation of $\mathbb{E}[F_{\mathrm{ST}}]$ by the leading eigenvalues of scaled PCA was less accurate than in analyses without admixed samples. In the CEU-ASW analysis for example, $\mathbb{E}[F_{\mathrm{ST}}]$ (= 4.55%) was lower than the leading eigenvalue of PCA (= 4.87%). An explanation for these discrepancies may be that F_{ST} is informative about the admixture proportion between admixed populations and their parental source populations [51]. With admixed samples, mismatches were also observed between the leading eigenvalue of the residual matrix and its prediction from RMT. The results suggest that the data do not agree with models of K discrete populations, and modified definitions of F_{ST} could be more appropriate for describing population structure in the presence of admixed individuals [52, 53].

Ancient DNA data

We illustrated how spectral estimates can be used to evaluate inbreeding coefficients from genotypes obtained after correction for experimental effects. We studied ancient DNA samples from early farmers from Anatolia (EFA, $n = 23$), steppe pastoralists from the Yamnaya culture (Steppe, $n = 15$), Western hunter-gatherers from Serbia (WHG, $n = 31$), and included Bell-Beaker samples from England and Germany (BKK, $n = 38$) [17, 54–56]. To estimate F_{ST} from those samples, we performed adjustment of pseudo-haploid genotypes for genomic coverage and for temporal distortions created by genetic drift. Temporal distortions were not expected to modify the average value of F_{ST} across loci. After genotypes were adjusted for coverage and corrected for distortions due to differences in sample ages, the resulting values could no longer be interpreted as allelic frequencies. The adjusted estimates for F_{ST} were equal to 4.7% for the *EFA—Steppe* data set, 5.8% for *EFA—WHG*, 5.1% for *Steppe—WHG* (Table 2). The separation condition was verified in all comparisons, and there was evidence of population structure in all pairwise analyses. Although individual PCA scores were impacted by coverage and temporal distortions (Fig 5), those unwanted effects did not generate substantial bias for PCA eigenvalues, leaving us with F_{ST} estimates that were similar with or without adjustment. For a three-population model including the EFA, WHG, and Steppe samples, the adjusted estimate for F_{ST} was equal to 7.0%, slightly lower than the uncorrected estimate (7.2%, Table 2) and than the sum of the leading values of PCA (equal to 7.3%). The smallest eigenvalue of the between-population matrix was larger (2.6%) than the leading eigenvalue of the residual matrix (1.8%). This was no longer the case when the Bell Beakers from England and Germany were included in the data set. With Bell Beaker samples, the smallest eigenvalue of the between-population matrix was lower (1.0%) than the leading eigenvalue of the residual

Table 2. F_{ST} estimates for ancient Eurasian samples with correction for genomic coverage.

	F_{ST} without correction	F_{ST} with correction	Lead. eigen. res. matrix*	RMT threshold**
EFA-Steppe	4.8%	4.7%	3.1%	2.8%
EFA-WHG	5.9%	5.8%	3.3%	2.0%
Steppe-WHG	5.2%	5.1%	3.8%	2.3%
EFA-Steppe-WHG	7.2%	7.0%	1.8% (2.6%)	1.5%
EFA-Steppe-WHG-BBK	5.9%	5.8%	1.8% (1.0%)	1.0%

EFA: Early Farmers from Anatolia, **WHG**: Western Hunter-Gatherers, **Steppe**: Yamnaya pastoralists, **BBK**: Bell Beakers from England and Germany.

* Leading eigenvalue of the within-population residual matrix (smallest eigenvalue of the between-population matrix)

** RMT threshold for evidence of population structure in pairs: $(1/\sqrt{L} + 1/\sqrt{n-1})^2$, RMT approximation in triplet and quadruplet: $(1 - \mathbb{E}[F_{\mathrm{ST}}])(1/\sqrt{L} + 1/\sqrt{n-K})^2$. L: number of loci, n: sample size.

https://doi.org/10.1371/journal.pgen.1009665.t002

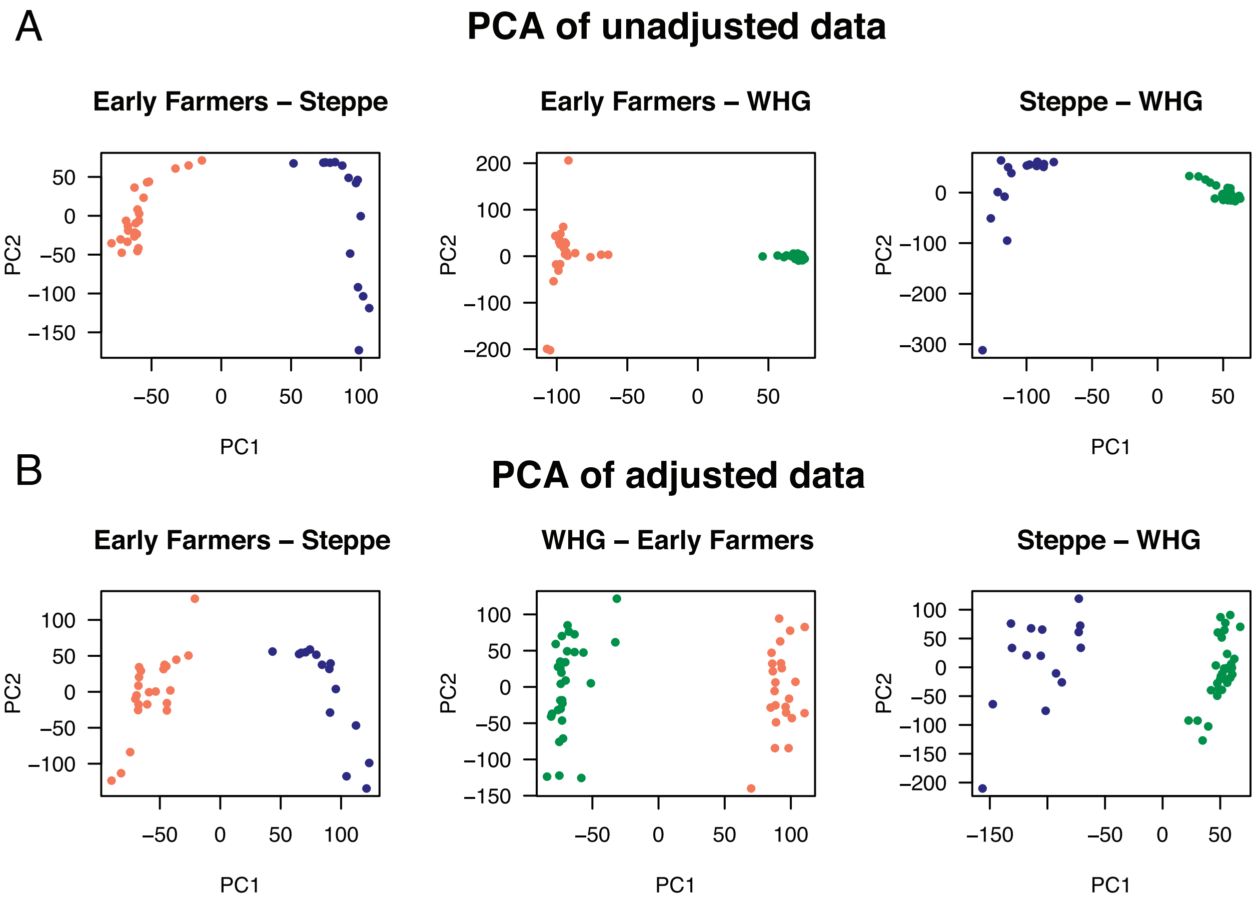

Fig 5. Correction for coverage in PC plots for pairs of ancient population samples. (**A**) PCA of unadjusted genotypes. (**B**) PCA of non-binary genotypic data adjusted for coverage. Population samples: Early Farmers (salmon color, $n_1 = 23$), Steppe pastoralists (blue color, $n_2 = 15$), (Western Hunter Gatherers, green color, $n_3 = 31$).

https://doi.org/10.1371/journal.pgen.1009665.g005

matrix (1.0%, Table 2). An explanation for this result is that the shared ancestry of Bell Beaker individuals [56] made PCA results inconsistent with a four-population model.

Genetic differentiation explained by bioclimatic factors

To provide a second illustration of the use of spectral estimates of inbreeding coefficients, we studied the role of bioclimatic factors in shaping population genetic structure in plants [29]. Here, the objective was to evaluate the proportion of differentiation explained by temperature and precipitation, which may influence adaptive genetic variation in those taxa. For 241 Swedish accessions of *Arabidopsis thaliana* taken from The 1,001 Genomes database [30], population structure was first evaluated by using a spatially explicit Bayesian algorithm. The individuals were clustered in two groups located in southern and northern Sweden (Fig 6A). For the groups estimated by spatial population structure analysis, the mean value of F_{ST} across loci was equal to 7.9%. This value was larger than the leading eigenvalue of the within-population matrix, equal to 4.9%. The proportion of variance explained by the first PCA axis was equal to 8.5%, greater than F_{ST} (Fig 6). An explanation for this discrepancy is that a two-population model may not fit the data accurately, as PCA axes can capture spatial genetic variation unseen by the discrete population model. Population structure was further evaluated by using $K = 3$ ancestral populations. Southern individuals were split into two groups along a East-West

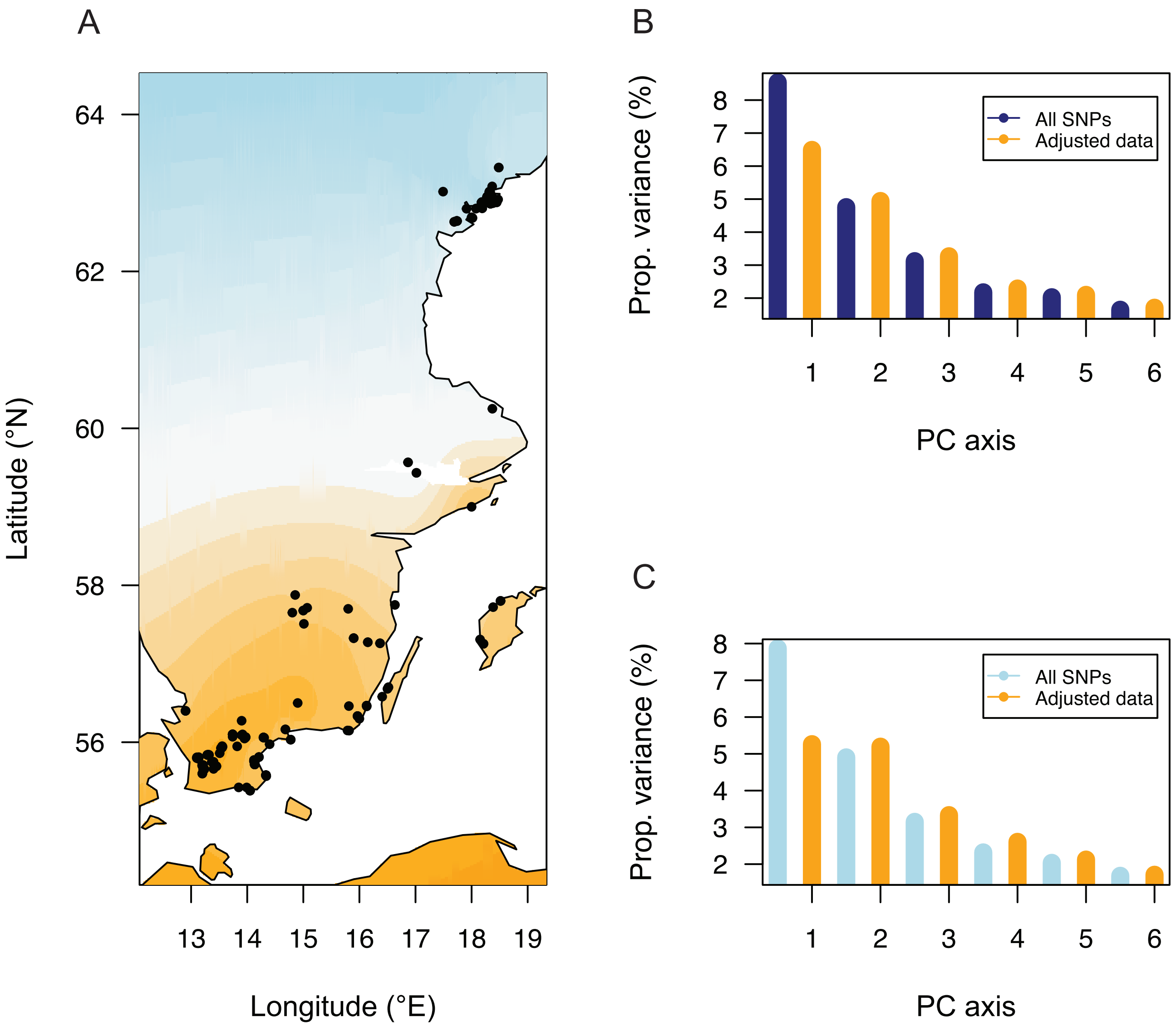

Fig 6. Neutral F_{ST} for *Arabidopsis thaliana* in Scandinavia. (**A**) Geographic locations of 241 samples and inference of population structure from a spatial method (Blue color: Southern cluster, Orange color: Northern cluster). The map was generated by using the open source R package maps (CC-BY license), which loads data from www.naturalearthdata.com. (**B**) Proportion of variance explained by PC axes before adjustment of genotypes for bioclimatic variables (blue color) and after adjustment (orange color). (**C**) Proportion of variance explained by the first axis of the between-population matrix, and by the first axes of the residual matrix (five components) before adjustment for bioclimatic variables (blue color) and after adjustment (orange color). Wright's coefficients are represented by the values for the first axis.

https://doi.org/10.1371/journal.pgen.1009665.g006

axis, exhibiting mixed ancestry from those groups (S10 Fig). With three groups, the leading eigenvalues of the between-population matrix were equal to 7.8% and 2.5%. The second eigenvalue was lower than the leading eigenvalue of the within-population matrix, equal to 3.7%. According to those results, we decided to focus on information carried by a two-population model.

After adjusting for bioclimatic variation, the leading eigenvalue of the PCA was equal to 6.5% (Fig 6C). The eigenvalue of the between-population matrix—which defines the average value of F_{ST} for the adjusted genotypes, was equal to $\mathbb{E}[F_{ST}^{adj}] = 5.3\%$. The second and subsequent PCA eigenvalues were equal to 4.9%, 3.2%, 2.3%, and those values were unaffected by the bioclimatic variables (Fig 6B). In addition, those eigenvalues agreed with the largest eigenvalues of the residual matrix, $Z_{\mathrm{S}}^{\mathrm{adj}}$, which were equal to 5.1%, 3.3%, 2.6% (Fig 6B). Comparing $\mathbb{E}[F_{\mathrm{ST}}^{\mathrm{adj}}]$ to $\mathbb{E}[F_{\mathrm{ST}}]$ [42], these results showed that the relative proportion of variation explained by climate along the first axis was around 33%. The results provide evidence that climate had an impact on the differentiation of populations along the south-north axis, but had less influence on other axes of genetic variation. In summary, those numbers suggest that bioclimatic conditions played a major role in driving genetic divergence between northern and southern populations of Scandinavian *A. thaliana*.

Conclusions

Assuming a model with K discrete populations, our study established a relationship between Wright's inbreeding coefficient, F_{ST}, and the $(K - 1)$ leading eigenvalues of the between-population matrix and of scaled PCA. A similar relationship was established between Nei's among-population diversity, D_{ST}, and the leading eigenvalues of unscaled PCA. Those relationships justify the use of PCA to describe population genetic structure from large genotype matrices. They extend results obtained from coalescent theory for two divergent populations in Ref. [16] to any discrete population model. Assuming that RMT holds for residual matrices, they increase the accuracy of previous results, clarifying for which sample sizes and number of loci they could be valid. Computing the eigenvalues of the between-population matrix and of the residual matrix can be done numerically with a computing cost similar to PCA (Fig 1). In our simulations, we found that the leading eigenvalue of the residual matrix was well predicted by RMT. RMT also provided a threshold value of F_{ST}, equal to $\theta = (1/\sqrt{L} + 1/\sqrt{n-1})^2$ below which there is no evidence of population structure for two or more populations. This threshold differs from the threshold value of $1/\sqrt{nL}$ proposed in Ref. [10], and it was supported by simulations of single population models. In addition to connecting the PCA of a genotype matrix to inbreeding coefficients and related quantities, our results have several implications for the analysis of adjusted genotypes, providing statistics analogous to F_{ST} and Q_{ST} for those data. Adjusted genotypes arise in many applications, such as ancient DNA, to correct for biases due to technical or sampling artifacts, or in ecological genomics where it allows evaluating the part of population differentiation explained by environmental variation. The proposed estimates of inbreeding coefficients are thus of great importance to the understanding of the demographic history of populations and their adaptation to environmental variation.

Methods

PCA and SVD

For a genotype matrix $\mathbf{X}$ with L loci, centered PCA computes the eigenvalues, $\sigma_i^2(\mathbf{Z})$, of the empirical covariance matrix, $\mathbf{Z}\mathbf{Z}^T/n$, where $\mathbf{Z} = \mathbf{Z}_c$ is the centered matrix, for which the mean value of each column has been substracted from $\mathbf{X}$ [9, 26]. Scaled PCA computes the eigenvalues, $\rho_i^2(\mathbf{Z}^{\mathrm{sc}})$, of the empirical correlation matrix, $\mathbf{Z}^{\mathrm{sc}}(\mathbf{Z}^{\mathrm{sc}})^T/n$, obtained for $\mathbf{Z}^{\mathrm{sc}}$, the matrix in which each column of $\mathbf{Z}$ is divided by $\sqrt{P(1-P)}$. In order to obtain unbiased estimates, empirical covariance and correlation matrices are usually divided by $(n - 1)$ instead of n. To avoid this complication, we kept n in all theoretical analyses (assuming that n is large). Unbiased estimates were used in empirical and simulated data analyses. Using the equivalence between PCA and SVD, the eigenvalues of PCA were computed as the squared non-null singular values of the matrix $\mathbf{Z}/\sqrt{n}$.

Spectral analysis

To make arguments easier to follow, we developed the analysis of eigenvalues for unscaled PCA. Extension to scaled PCA does not create mathematical complications but has heavier notations. This paragraph sketches the key arguments for the main result. More details are provided in S1 Text. We found that the squared Hilbert-Schmidt norm of the between-population matrix $\mathbf{Z}_{\mathrm{ST}}$ is equal to

$$\|\mathbf{Z}_{\mathrm{ST}}\|^2 = nL\,\mathbb{E}\left[\sum_{k=1}^{K} c_k \left(\sum_{j=1}^{K} c_j (p_j - p_k)\right)^2\right] = nL \times \mathbb{E}[D_{\mathrm{ST}}]/2\,,$$

where the mathematical symbol $\mathbb{E}[Q]$ denotes the mean value of a quantitative or discrete quantity Q over the L loci. For scaled PCA, the squared norm is equal to $nL \times \mathbb{E}[F_{\mathrm{ST}}]$. The matrices $\mathbf{Z}_{\mathrm{ST}}$ and $\mathbf{Z}_{\mathrm{S}}$ satisfy orthogonality conditions. In particular, the matrices are related by the Pythagorean formula $\|\mathbf{Z}\|^2 = \|\mathbf{Z}_{\mathrm{ST}}\|^2 + \|\mathbf{Z}_{\mathrm{S}}\|^2$. The Pythagorean formula is the main argument for the alternative form of Theorem 1. In addition, stronger orthogonality conditions hold and enable showing that the accuracy of approximation of eigenvalues is of order $1/L$.

F-models

F-models are models for K discrete populations diverging from an ancestral gene pool [33]. In the ancestral gene pool, the derived allele is present with frequency p_{anc}. The K populations diverged from each other and from the ancestral population with drift coefficient equal to F_k relative to the ancestral pool. Conditional on p_{anc}, the allele frequency at a particular locus in population k follows a beta distribution of shape parameters $p_{\mathrm{anc}}(1 - F_k)/F_k$ and $(1 - p_{\mathrm{anc}})(1 - F_k)/F_k$. To create a distribution over the L loci, p_{anc} is drawn from a beta distribution with shape parameters a and b, leading to $\mathbb{E}[p_{\mathrm{anc}}] = a/(a + b)$. The expected ancestral heterozygozity, H_A, is equal to $\mathbb{E}[H_{\mathrm{A}}] = 2ab/(a + b)(a + b + 1)$. For F-models, the expected value of D_{ST} can be formulated as $\mathbb{E}[D_{\mathrm{ST}}] = \mathbb{E}[H_{\mathrm{A}}] \sum_{k=1}^{K} c_k(1 - c_k)F_k$ (S1 Text). Numerical values for $\mathbb{E}[F_{\mathrm{ST}}]$ are less explicit, but they can be obtained by using Monte-Carlo simulations.

Simulations of F-models were performed in the R programming language. We performed simulations of single population models ($K = 1$) to check whether approximations derived from RMT appropriately describe the leading eigenvalue of scaled PCA in the absence of population structure. Simulations of F-models were performed with a value of the drift coefficient equal to $F = 15\%$. The ancestral frequency for the derived allele, p_{anc}, was drawn from a beta distribution with shape parameters $a = 1$ and $b = 9$, so that $\mathbb{E}[p_{\mathrm{anc}}] = 10\%$ (S1 Fig). Simulations of F-models were performed with $K = 2$ to check whether the data could fit theoretical expectations for D_{ST} and F_{ST}. Two hundred simulations of F-models were performed with equal values of the drift coefficients randomly drawn between 1% and 75% ($F_1 = F_2$). The ancestral frequency for the derived allele, p_{anc}, was drawn from a beta distribution with shape parameters $a = 1$ and $b = 4$, so that $\mathbb{E}[p_{\mathrm{anc}}] = 20\%$. The total sample size was equal to $n = 100$ and the sample proportion c_1 was drawn from a uniform distribution between 10% and 50%. Next, we considered three-population F-models with equal sample sizes and ancestral allele frequencies distributed according to the uniform distribution, ($a = 1$ and $b = 1$). With the uniform distribution, we found that $\mathbb{E}[H_{\mathrm{A}}] = 1/3$, and the non-null eigenvalues of the between-population covariance matrix could be computed as $\lambda_i = (F_1 + F_2 + F_3 \pm \sqrt{F_1^2 + F_2^2 + F_3^2 - F_1F_2 - F_2F_3 - F_3F_1})/54$, for $i = 1, 2$ (S1 Text). We had $\mathbb{E}[D_{\mathrm{ST}}] = 2(\lambda_1 + \lambda_2)$. Two hundred simulations of three-population models were performed with unequal drift coefficients drawn between 1% and 25%. The total sample size was equal to $n = 100$ and the number of loci was equal to $L = 20,000$. For values of n between 30 and 300, and number of loci between 100 and 1000, we performed additional simulations with small drift coefficients ($F_k \leq 10\%$) to evaluate the probability that the data verify the separation condition. We also performed simulations of F-models for K between 4 and 40, using $n_k = 20$ individuals in each sample ($L = 20,000$) and two models of drift: $F_k = 0.1$ and $F_k = 0.2/k$ for all $k = 4, \ldots, 40$.

Approximation of the residual matrix from Random Matrix Theory (RMT)

For discrete population models, approximations of eigenvalues for the within-population (residual) matrix were obtained from RMT [10, 57–60]. RMT considers large sample sizes, and keeps the ratio of the number of loci to the sample size at a constant value, $\gamma = L/n$. For single population models, we have $\mathbf{Z} = \mathbf{Z}_S$, and the proportions of variance explained by each principal axis were approximated by the Marchenko-Pastur probability density function described by

$$p(x) = L\frac{\sqrt{(x_M - x)(x - x_m)}}{2x\pi}\,, \quad x_m = (1-\sqrt{\gamma})^2/L \leq x \leq x_M = (1+\sqrt{\gamma})^2/L\,,$$

and the proportion of variance explained by the first principal axis was approximated by $(1/\sqrt{L} + 1/\sqrt{n-1})^2$. For $K > 1$, the Marchenko-Pastur density modelled the bulk distribution of eigenvalues for the within-population (residual) matrix. Under the separation condition Eq (7), the proportion of variance explained by the Kth principal axis was approximated by $(1 - \mathbb{E}[F_{ST}])(1/\sqrt{L} + 1/\sqrt{n-K})^2$. Regarding unscaled PCA, the largest singular value of the within-population matrix was approximated by $2\sigma_1^2(\mathbf{Z}_S)/L \approx \mathbb{E}[H_S](1/\sqrt{L} + 1/\sqrt{n-K})^2$. If there truly is a single population represented in the total sample, then F_{ST} for two equal size samples should be of order $(1/\sqrt{L} + 1/\sqrt{n-1})^2$.

Human DNA analyses

We computed PCA eigenvalues, the mean values of F_{ST} and the leading eigenvalues of the residual matrix for pairs and triplets including human population samples from The 1000 Genomes Project [50]. In these comparisons, the number of SNPs was $L \approx 1.3$M, after filtering out for minor allele frequency less than 5%. We considered samples from Han Chinese in Beijing (CHB, $n = 100$), Yoruba (YRI, $n = 158$), Utah residents with European ancestry (CEU, $n = 104$), Iberian (IBS, $n = 147$). We considered samples from populations with mixed ancestry, Americans of African Ancestry in SW USA (ASW, $n = 97$), Colombians from Medellin Colombia (CLM, $n = 102$), Puerto Ricans from Puerto Rico (PUR, $n = 94$), individuals of Mexican Ancestry from Los Angeles USA (MXL, $n = 100$), and African Caribbeans in Barbados (ACB, $n = 98$). Some pairs and triplets included individuals with mixed ancestry.

Ancient DNA analyses

We analyzed 143,081 pseudo-haploid SNP genotypes from ancient samples of early farmers from Anatolia ($n = 23$), steppe pastoralists from the Yamnaya culture ($n = 15$), Western hunter-gatherers from Serbia ($n = 31$), and Bell-Beakers from England and Germany ($n = 38$). The data were extracted from a public data set available from David Reich lab's repository (reich.hms.harvard.edu) [17, 54–56]. The ancient samples had a minimum coverage of 0.25x, a median coverage of 2.69x (mean of 2.98x) and a maximum coverage of 13.54x. Genotypes were adjusted for coverage by fitting a latent factor regression model with the number of factors equal to the number of sample minus two. The matrix was then adjusted for distortions due to differences in sample ages, resulting in surrogate genotypes encoded as continuous variables without any direct interpretation in terms of allelic frequency [28].

Genomic and bioclimatic data analyses

We studied 241 swedish plant accessions from The 1,001 Genomes database for *Arabidopsis thaliana* [30]. A matrix of SNP genotypes was obtained by considering variants with minor allele frequency greater than 5% and a density of variants around one SNP every 1,000 bp (167,475 SNPs). The individuals were clustered in groups based on analysis of population structure accounting for geographic proximity [61]. Global climate and weather data corresponding to individual geographic coordinates were downloaded from the WorldClim database (https://worldclim.org). Eighteen bioclimatic variables, derived from the monthly temperature and rainfall values, were considered as representing the current environmental matrix. Correction of genotypes for locus-specific effects of the eighteen environmental variables was performed with a latent factor regression model implemented in the R package `lfmm` [41]. For the matrix of centered genotypes, **Z**, and the matrix of eighteen bioclimatic variables, **Y**, the program estimated a matrix of surrogate genotypes, **W**, by adjusting a regression model of the form $\mathbf{Z} = \mathbf{Y}\mathbf{B}^T + \mathbf{W} + \epsilon$. To keep the latent matrix estimate (**W**) as close as possible to **Z**, we used $k = n - 19 = 222$ factors to compute **W**. The codes necessary to reproduce the data analyses presented in this study are available in an `R` package at https://github.com/bcm-uga/spectralfst.

Supporting information

S1 Text. Mathematical details for the main results and for results on *F*-models.
(PDF)

S1 Table. F_{ST} estimates for populations from The 1,000 Genomes Project.
(PDF)

S1 Fig. Random matrix theory approximation of proportions of variance in single population *F*-models. Two simulations of *F*-models were performed with p_{anc} drawn from a beta distribution with shape parameters $a = 1$ and $b = 9$, and $F = 15\%$. **Top row**: $n = 200$ individuals and $L = 69, 248$ SNPs. **Bottom row**: $n = 50$ individuals and $L = 10, 331$ SNPs. **SFS**: Site Frequency Spectrum, **MP approximation**: Marchenko-Pastur approximation of the distribution of scaled PCA eigenvalues (blue curve). Histograms of scaled PCA eigenvalues representing the proportions of variance explained by the $(n - 1)$ principal axes are displayed in grey color.
(PDF)

S2 Fig. Separation of variance components in artificial population samples built from single population *F*-models. For each value of the drift coefficient, each couple of blue and orange dots represent a simulated data set. F_{ST} (blue dots) corresponds to the non-null eigenvalue of the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}$, and the "residual" value corresponds to the leading eigenvalue of $\mathbf{Z}_{\mathrm{S}}$. Population structure is detected when the blue dot is above the orange dot. **Top row**: 200 simulations with $n = 100$ individuals and L around 10,000 SNPs. **Bottom row**: 200 simulations with $n = 10$ individuals and L around 1,000 SNPs. Simulations were performed with ancestral frequencies, p_{anc}, drawn from a beta distribution with shape parameters $a = 1$ and $b = 9$.
(PDF)

S3 Fig. Accuracy of D_{ST} estimates with respect to their theoretical values in two-population *F*-models. Comparison of values of D_{ST} averaged over loci and their theoretical values in *F*-models. Simulations of *F*-models were performed with equal drift coefficients ($F_1 = F_2$) ranging between 1% and 75%, and sample proportions c_1 between 10% and 50% ($n = 100$

individuals). The ancestral frequencies, p_{anc}, were drawn from a beta distribution with shape parameters $a = 1$ and $b = 4$.
(PDF)

S4 Fig. Accuracy of approximation of F_{ST} and separation condition in two-population *F*-models. First column: Approximation error defined as the difference between F_{ST} and the leading eigenvalue of scaled PCA, $\mathbb{E}[F_{\text{ST}}] - \rho_1^2(\mathbf{Z}^{\text{sc}})/L$, as a function of the difference of eigenvalues, $\rho_1^2(\mathbf{Z}^{\text{sc}}_{\text{ST}})/L - \rho_1^2(\mathbf{Z}^{\text{sc}}_{\text{S}})/L$. **Second column**: Approximation errors according to whether the separation condition is checked or not. **Third column**: Leading eigenvalue of scaled PCA as a function of $\mathbb{E}[F_{\text{ST}}]$. Simulations of *F*-models were performed for n individuals and L loci with equal drift coefficients $F_1 = F_2 = 0.02$. Ancestral frequencies, p_{anc}, were drawn from a beta distribution with shape parameters $a = 1$ and $b = 4$.
(PDF)

S5 Fig. Approximation errors decrease as $1/L$ in two-population *F*-models. Approximation error defined as the absolute difference between F_{ST} and the leading eigenvalue of scaled PCA, $\mathbb{E}[F_{\text{ST}}] - \rho_1^2(\mathbf{Z}^{\text{sc}})/L$ as a function of $1/L$ (L is the number of unlinked loci). The red line corresponds to the linear regression Log(Approx) = $a + b$ Log(L), and has slope equal to $b = -1.001$ ($R^2 = 0.997$, $P < 2\text{e-}16$). Simulations of *F*-models were performed for $n = 150$ individuals with drift coefficients equal to $F_1 = F_2 = 0.02$. The ancestral frequencies, p_{anc}, were drawn from a beta distribution with shape parameters $a = 1$ and $b = 4$.
(PDF)

S6 Fig. First eigenvalue (F_{ST} estimate) as a function of sample size in two-population *F*-models. (**A**) $L = 100$ loci: The separation condition was verified for sample sizes > 60. (**B**) $L = 100,000$ loci: The separation condition was verified for all sample sizes. F_{ST}: Leading eigenvalue of the PCA. **RMT threshold**: Approximation of the detection threshold from RMT, equal to $(1/\sqrt{L} + 1/\sqrt{n-1})^2$. Dashed line: Theoretical value for an infinite sample size, $\mathbb{E}[F_{\text{ST}}] = 3.97\%$. Simulations of *F*-models were performed with ancestral frequencies drawn from a beta(1,4) distribution and with $F_1 = F_2 = 10\%$.
(PDF)

S7 Fig. Leading eigenvalues in three-population *F*-models. (**A**) Sum of the first two eigenvalues of centered PCA as a function of the mean of $D_{\text{ST}}/2$ across loci. (**B**) First eigenvalue of centered PCA as a function of its expected value $\lambda_1 = (F_1 + F_2 + F_3 + \sqrt{F_1^2 + F_2^2 + F_3^2 - F_1F_2 - F_2F_3 - F_3F_1})/54$. (**C**) Sum of the first two eigenvalues of scaled PCA as a function of the mean of F_{ST} across loci. (**D**) Third eigenvalue of scaled PCA as a function of its approximation from RMT. **MP approximation**: Marchenko-Pastur approximation of the largest eigenvalue of the residual matrice, $\mathbf{Z}_{\text{S}}/\sqrt{n-3}$, equal to $(1 - \rho_1^2 - \rho_2^2) \times (1/\sqrt{L} + 1/\sqrt{n-3})^2$. The dashed lines correspond to the diagonal $y = x$. Simulations of *F*-models were performed for $n = 100$ individuals with drift coefficients F_1, F_2, F_3 between 1% and 25%, equally sampled populations, and ancestral frequencies drawn from the uniform distribution ($L = 20000$ loci).
(PDF)

S8 Fig. Separation condition in three-population *F*-models. Probability that the separation condition is verified for sample sizes ranging between $n = 30$ and $n = 300$ individuals, and number of loci ranging between $L = 100$ and $L = 1000$. Simulations of *F*-models were performed with equal sample sizes, random drift coefficients lower than 10%, and ancestral frequencies drawn from the uniform distribution. Five hundred simulations were performed for

each combination of n and L.
(PDF)

S9 Fig. Accuracy of approximation of F_{ST} by PCA for K-population F-models. A-B) Simulations with equal drift coefficients $F_k = 0.1$. The accuracy of the approximation of $\mathbb{E}[F_{ST}]$ by the sum of the $K-1$ leading eigenvalues of the PCA is comparable to K/L (A) and the separation of eigenvalues from the residual matrix is verified (B). **C-D**) Simulations with unequal drift coefficients $F_k = 0.2/k$. The accuracy of the approximation of $\mathbb{E}[F_{ST}]$ diverged from K/L when the separation of eigenvalues did not hold. In all simulations, ancestral frequencies were drawn from a beta distribution with shape parameters $a = 1$ and $b = 4$. Sub-population samples had size equal to $n_k = 20$, the total population size was $n = 20 \times K$), and the number of SNP loci was $L \approx 19850$.
(PDF)

S10 Fig. Estimates of ancestry coefficients for 241 Swedish accessions of *A. thaliana*. Ancestry coefficients obtained from the spatially explicit ancestry estimation program `tess3r` with $K = 3$ populations. The southern group exhibits substantial levels of mixed ancestry.
(PDF)

Author Contributions

Conceptualization: Olivier François, Clément Gain.

Formal analysis: Olivier François, Clément Gain.

Funding acquisition: Olivier François.

Investigation: Olivier François.

Methodology: Olivier François, Clément Gain.

Project administration: Olivier François.

Software: Clément Gain.

Supervision: Olivier François.

Validation: Olivier François, Clément Gain.

Writing – original draft: Olivier François.

Writing – review & editing: Olivier François, Clément Gain.

References

1. Wright S. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution. 1965; 19:395–420. https://doi.org/10.2307/2406450
2. Malécot G. Les mathématiques de hérédité. Paris: Masson; 1948.
3. Cockerham CC. Variance of gene frequencies. Evolution. 1969; 23:72–84. https://doi.org/10.2307/2406485 PMID: 28562963
4. Nei M. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA. 1973; 70:3321–3323. https://doi.org/10.1073/pnas.70.12.3321 PMID: 4519626
5. Weir BS, Cockerham CC. Estimating *F*-statistics for the analysis of population structure. Evolution. 1984; 38:1358–1370. https://doi.org/10.2307/2408641 PMID: 28563791
6. Slatkin M. Inbreeding coefficients and coalescence times. Genet Res. 1991; 58:67–175. https://doi.org/10.1017/S0016672300029827 PMID: 1765264
7. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet. 2009; 10:639–650. https://doi.org/10.1038/nrg2611 PMID: 19687804

8. Hotelling H. Relations between two sets of variates. Biometrika. 1936; 28:321–377. https://doi.org/10.1093/biomet/28.3-4.321

9. Jolliffe I. Principal component analysis. New York: Springer-Verlag; 1986.

10. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2: e0020190. https://doi.org/10.1371/journal.pgen.0020190 PMID: 17194218

11. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. https://doi.org/10.1093/genetics/155.2.945 PMID: 10835412

12. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164:1567–1587. https://doi.org/10.1093/genetics/164.4.1567 PMID: 12930761

13. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Phil Trans R Soc A. 2016; 374:20150202. https://doi.org/10.1098/rsta.2015.0202 PMID: 26953178

14. Cavalli-Sforza LL, Edwards AWF, Geerts S. Analysis of human evolution. In: Genetics today: Proceedings of the 11th International Congress of Genetics, The Hague, The Netherlands. New York: Pergamon. 3:923-993;1963.

15. Menozzi P, Piazza A, Cavalli-Sforza LL. Synthetic maps of human gene frequencies in Europeans. Science. 1978; 201:786–792. https://doi.org/10.1126/science.356262 PMID: 356262

16. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009; 5: e1000686. https://doi.org/10.1371/journal.pgen.1000686 PMID: 19834557

17. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015; 522:207. https://doi.org/10.1038/nature14317 PMID: 25731166

18. Zheng X, Weir BS. Eigenanalysis of SNP data with an identity by descent interpretation. Theor Pop Biol. 2016; 107:65–76. https://doi.org/10.1016/j.tpb.2015.09.004 PMID: 26482676

19. Bryc K, Bryc W, Silverstein JW. Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. Theor Pop Biol. 2013; 89:34–43. https://doi.org/10.1016/j.tpb.2013.08.004 PMID: 23973732

20. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol. 2016; 33:1082–1093. https://doi.org/10.1093/molbev/msv334 PMID: 26715629

21. Chen GB, Lee SH, Zhu ZX, Benyamin B, Robinson MR. EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. Heredity. 2016; 117:51–61. https://doi.org/10.1038/hdy.2016.25 PMID: 27142779

22. Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson N, Price AL. Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. Am J Hum Genet. 2016; 98:456–472. https://doi.org/10.1016/j.ajhg.2015.12.022 PMID: 26924531

23. François O, Martins H, Caye K, Schoville SD. Controlling false discoveries in genome scans for selection. Mol Ecol. 2016; 25:454–469. https://doi.org/10.1111/mec.13513 PMID: 26671840

24. Wilkinson-Herbots HM. Genealogy and subpopulation differentiation under various models of population structure. J Math Biol. 1998; 37:535–585. https://doi.org/10.1007/s002850050140

25. Ma J, Amos CI. Theoretical formulation of principal components analysis to detect and correct for population stratification. PLoS ONE. 2010; 5:e12510. https://doi.org/10.1371/journal.pone.0012510 PMID: 20862251

26. Johnstone IM, Paul D. PCA in high dimensions: An orientation. Proc IEEE. 2018; 106:1277–1292. https://doi.org/10.1109/JPROC.2018.2846730 PMID: 30287970

27. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014; 15:356. https://doi.org/10.1186/s12859-014-0356-4 PMID: 25420514

28. François O, Jay F. Factor analysis of ancient population genomic samples. Nat Commun. 2020; 11:4661 https://doi.org/10.1038/s41467-020-18335-6 PMID: 32938925

29. Wang IJ, Glor RE, Losos JB. Quantifying the roles of ecology and geography in spatial genetic divergence. Ecol Lett. 2013; 16:175–182. https://doi.org/10.1111/ele.12025 PMID: 23137142

30. The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016; 166:481–491. https://doi.org/10.1016/j.cell.2016.05.063 PMID: 27293186

31. Li Z, Löytynoja A, Fraimout A, Merilä J. Effects of marker type and filtering criteria on $Q_{ST} - F_{ST}$ comparisons. Royal Soc Open Sci. 2019; 6:190666. https://doi.org/10.1098/rsos.190666

32. Wright S. Evolution in Mendelian populations. Genetics. 1931; 16:97–159. https://doi.org/10.1093/genetics/16.2.97 PMID: 17246615

33. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995; 96:3–12. https://doi.org/10.1007/BF01441146 PMID: 7607457

34. Nei M, Chesser RK. Estimation of fixation indices and gene diversities. Ann Hum Genet. 1983; 47:253–259. https://doi.org/10.1111/j.1469-1809.1983.tb00993.x PMID: 6614868

35. Culley TM, Wallace LE, Gengler-Nowak KM, Crawford DJ. A comparison of two methods of calculating G_{ST}, a genetic measure of population differentiation. Am J Bot. 2002; 89:460–465. https://doi.org/10.3732/ajb.89.3.460 PMID: 21665642

36. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. 2013; 23:1514–1521. https://doi.org/10.1101/gr.154831.113 PMID: 23861382

37. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006; 7(10):781–791. https://doi.org/10.1038/nrg1916 PMID: 16983374

38. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013; 14(5):1–20. https://doi.org/10.1186/gb-2013-14-5-r51 PMID: 23718773

39. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28(6):882–883. https://doi.org/10.1093/bioinformatics/bts034 PMID: 22257669

40. Wang J, Zhao Q, Hastie T, Owen AB. Confounder adjustment in multiple testing. Ann Stat. 2017; 45(5):1863–1894. https://doi.org/10.1214/16-AOS1511 PMID: 31439967

41. Caye K, Jumentier B, Lepeule J, François O. LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. Mol Biol Evol. 2019; 36:852–860. https://doi.org/10.1093/molbev/msz008 PMID: 30657943

42. Spitze K. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. Genetics. 1993; 135:367–374. https://doi.org/10.1093/genetics/135.2.367 PMID: 8244001

43. Whitlock MC. Evolutionary inference from Q_{ST}. Mol Ecol. 2008; 17(8):1885–1896. https://doi.org/10.1111/j.1365-294X.2008.03712.x PMID: 18363667

44. Linck E, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Res. 2019; 19:639–647. https://doi.org/10.1111/1755-0998.12995 PMID: 30659755

45. Cattell RB. The scree test for the number of factors. Multivariate Behav Res. 1966; 1:245–276. https://doi.org/10.1207/s15327906mbr0102_10 PMID: 26828106

46. Bookstein FL. Pathologies of between-groups principal components analysis in geometric morphometrics. Evol Biol. 2019; 46:271–302. https://doi.org/10.1007/s11692-019-09484-8

47. Cardini A, O'Higgins P, Rohlf FJ. Seeing distinct groups where there are none: spurious patterns from between-group PCA. Evol Biol. 2019; 46:303–316. https://doi.org/10.1007/s11692-019-09487-5

48. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192:1065–1093. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212

49. Peter BM. Admixture, population structure, and *F*-statistics. Genetics. 2016; 202:1485–1501. https://doi.org/10.1534/genetics.115.183913 PMID: 26857625

50. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. https://doi.org/10.1038/nature15393

51. Boca SM, Rosenberg NA. Mathematical properties of F_{ST} between admixed populations and their parental source populations. Theor Popul Biol. 2011; 80:208–216. https://doi.org/10.1016/j.tpb.2011.05.003 PMID: 21640742

52. Martins H, Caye K, Luu K, Blum MGB, François O. Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. Mol Ecol. 2016; 25:5029–5042. https://doi.org/10.1111/mec.13822 PMID: 27565448

53. Ochoa A, Storey JD. Estimating F_{ST} and kinship for arbitrary population structures. PLoS Genet. 2021; 17:e1009241. https://doi.org/10.1371/journal.pgen.1009241 PMID: 33465078

54. Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015; 522:167–172. https://doi.org/10.1038/nature14507 PMID: 26062507

55. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015; 528:499. https://doi.org/10.1038/nature16152 PMID: 26595274

56. Mathieson I, Roodenberg SA, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature. 2018; 555:197. https://doi.org/10.1038/nature25778 PMID: 29466330

57. Marčenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. Mat Sb. 1967; 1:457. https://doi.org/10.1070/SM1967v001n04ABEH001994

58. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Ann Stat. 2001; 29:295–327. https://doi.org/10.1214/aos/1009210543

59. Johnstone IM. Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy-Widom limits and rates of convergence. Ann Stat. 2008; 36:2638–2716 PMID: 20157626

60. Bryson J, Vershynin R, Zhao H. Marchenko-Pastur law with relaxed independence conditions. arXiv:1912.12724 [Preprint]. 2019. Available from: https://arxiv.org/abs/1912.12724

61. Caye K, Jay F, Michel O, François O. Fast inference of individual admixture coefficients using geographic data. Ann Appl Stat. 2018; 12:586–608. https://doi.org/10.1214/17-AOAS1106